

Amendments to the Claims:

This listing of claims will replace all prior versions, and listings, of claims in the application:

Listing of Claims:

1. (Currently amended) A method for generating a representation of a document comprising:

 ~~sampling the document to obtain~~ obtaining a plurality of overlapping blocks by
sampling the document;

 choosing a subset of the plurality of overlapping blocks, where the subset is less
than an entirety of the plurality of overlapping blocks; and

 compacting the subset of the plurality of overlapping blocks to obtain the representation of the document.
2. (Previously presented) The method of claim 1, wherein compacting the subset of the plurality of overlapping blocks includes setting bits in the representation of the document based on the subset of the plurality of overlapping blocks.
3. (Original) The method of claim 1, wherein the representation of the document includes a fingerprint of a predetermined length.
4. (Original) The method of claim 3, wherein the predetermined length is eight or sixteen bytes.

5. (Previously presented) The method of claim 1, further comprising:
generating checksum values for the plurality of overlapping blocks.
6. (Previously presented) The method of claim 5, wherein choosing a subset of the plurality of overlapping blocks includes selecting a predetermined number of the smallest checksum values.
7. (Previously presented) The method of claim 5, wherein choosing a subset of the plurality of overlapping blocks includes selecting a predetermined number of the largest checksum values.
8. (Previously presented) The method of claim 1, further comprising:
hashing the subset of the plurality of overlapping blocks to a length for indexing the representation of the document.
9. (Previously presented) The method of claim 8, wherein hashing the subset of the plurality of overlapping blocks includes taking a number of least significant bits of the subset of the plurality of overlapping blocks.
10. (Previously presented) The method of claim 2, wherein setting the bits includes flipping a bit in the representation of the document when the bit corresponds to a block in the subset of plurality of overlapping blocks.

11. (Previously presented) The method of claim 1, wherein each of the plurality of overlapping blocks is of a predetermined length.

12. (Currently amended) The method of claim 11, wherein ~~sampling the document~~ obtaining a plurality of overlapping blocks further includes:

padding null characters to the document when a length of the document is below the predetermined length.

13. (Previously presented) A method for generating a representation of a document comprising:

sampling the document to obtain a plurality of overlapping samples;

selecting a predetermined number of the plurality of overlapping samples as those of the samples corresponding to a predetermined number of smallest samples or a predetermined number of largest samples; and

setting bits in the representation of the document based on the selected predetermined number of the samples.

14. (Previously presented) The method of claim 13, wherein the representation of the document includes a fingerprint of a predetermined length.

15. (Original) The method of claim 14, wherein the predetermined length is eight or sixteen bytes.

16. (Original) The method of claim 13, further comprising:
generating checksum values for the samples; and
selecting the predetermined number of the samples as those of the samples
corresponding to a predetermined number of the smallest checksum values or a
predetermined number of the largest checksum values.
17. (Original) The method of claim 13, further comprising:
hashing the predetermined number of the samples to a length for indexing the
representation of the document.
18. (Original) The method of claim 17, wherein hashing the predetermined
number of the samples includes taking a number of least significant bits of the
predetermined number of samples.
19. (Original) The method of claim 17, wherein setting bits in the
representation of the document includes flipping a bit in the representation of the
document when the bit is addressed by the hashed samples.
20. (Currently amended) A computer-implemented device comprising:
a memory to store instructions for implementing:
a fingerprint creation component to generate a fingerprint of a
predetermined length for an input document, the fingerprint generated by

sampling the input document to obtain samples,
choosing a subset of the samples, where the subset is less than an entirety of the samples, and
generating the fingerprint from the subset of the samples by
compacting the subset of the samples_{[[;]]}, and
a similarity detection component to compare pairs of fingerprints to
determine whether the pairs of fingerprints correspond to near-duplicate documents; and
a processor to execute the instructions in the memory.

21. (Previously presented) The computer-implemented device of claim 20,
further including:

a search engine to return documents to a user as a single link when the documents
are determined to correspond to near-duplicate documents.

22. (Previously presented) The computer-implemented device of claim 20,
wherein the similarity detection component compares the pairs of fingerprints by
calculating a hamming distance.

23. (Previously presented) The computer-implemented device of claim 22,
wherein the similarity detection component determines that the pairs of documents
correspond to near-duplicate documents when the hamming distance is below a threshold.

24. (Previously presented) The computer-implemented device of claim 20,

wherein the fingerprint creation component additionally:

chooses the subset as a predetermined number of smallest checksums calculated from the subset of the samples.

25. (Previously presented) The computer-implemented device of claim 20, wherein the fingerprint creation component additionally:

chooses the subset as a predetermined number of largest checksums calculated from the subset of the samples.

26. (Currently amended) A computer-implemented device comprising:
memory to store instructions for implementing:

means for sampling a document to obtain a plurality of overlapping blocks[[]],

means for choosing a subset of the plurality of overlapping blocks, where the subset is less than an entirety of the plurality of overlapping blocks[[]], and

means for compacting the subset of the plurality of overlapping blocks to obtain a compact representation of the document; and

a processor to execute the instructions in memory.

27. (Previously presented) The computer-implemented device of claim 26, further comprising:

means for calculating checksum values for the plurality of overlapping blocks, wherein the means for choosing a subset of the plurality of overlapping blocks chooses

the subset based on the checksum values.

28. (Previously presented) The computer-implemented device of claim 27, wherein the means for choosing a subset of the plurality of overlapping blocks chooses the subset as a predetermined number of the smallest checksum values.

29. (Previously presented) The computer-implemented device of claim 27, wherein the means for choosing a subset of the plurality of overlapping blocks chooses the subset as a predetermined number of the largest checksum values.

30. (Previously presented) The computer-implemented device of claim 27, wherein the means for compacting the subset of plurality of overlapping blocks includes means for flipping bits in the compact representation that are addressed by a hashed version of the checksum values.

31. (Currently amended) A computer-readable memory device containing program instructions that, when executed by a processor, cause the processor to:

sample a document to obtain a plurality of overlapping samples;

select a predetermined number of the plurality of overlapping samples as those of the samples corresponding to a predetermined number of [[the]] smallest samples or a predetermined number of largest samples; and

set bits in a representation of the document based on the selected predetermined number of the samples.

32. (Previously presented) The computer-readable memory device of claim 31, further including program instructions that, when executed by the processor, cause the processor to:

hash the predetermined number of the samples to a length for indexing the representation of the document.

33. (Previously presented) The computer-readable memory device of claim 32, wherein hashing the predetermined number of the samples includes taking a number of least significant bits of the predetermined number of samples.

34. (Currently amended) A computer-implemented device comprising:
memory to store instructions for implementing:
means for sampling a document to obtain a plurality of overlapping
blocks[[]],
means for calculating checksum values for the plurality of overlapping
blocks[[]],
means for choosing a subset of the plurality of overlapping blocks based
on the calculated checksum values, where the subset is less than an entirety of the
overlapping blocks[[]], and
means for setting bits in a compact representation of the document based
on the subset of the plurality of overlapping blocks for flipping bits in the compact
representation that are addressed by a hashed version of the checksum values; and

a processor to execute the instructions in memory.